Clustering Report for Customer Segmentation

1. Introduction

In this project, we performed customer segmentation using K-means clustering based on transaction data. The goal is to group customers into distinct clusters that share similar purchasing behavior, helping to tailor marketing strategies, promotions, and improve customer service.

2. Data Preprocessing and Feature Engineering

We began by aggregating transaction data at the customer level. The features used for clustering include:

- **TotalValue**: The sum of transaction values for each customer.
- **Quantity**: The total quantity of items purchased by each customer.
- **TransactionFrequency**: The count of transactions per customer (frequency of transactions).

Additionally, customer profile information such as **Region** was merged with the transaction data. The data was preprocessed as follows:

- **Normalization**: Numerical features (TotalValue, Quantity, TransactionFrequency) were normalized using **StandardScaler** to bring them to a similar scale.
- **Encoding Categorical Features**: The **Region** variable was encoded using one-hot encoding to create binary columns, excluding the first category to avoid multicollinearity.
- **Drop CustomerID**: The CustomerID column was dropped as it was not needed for the clustering process.

The final dataset contains the following features: TotalValue, Quantity, TransactionFrequency, and region-based dummy variables (e.g., Region_West, Region_East).

3. Clustering with K-Means

To determine the optimal number of clusters for segmentation, we used the **Elbow Method**. The method involved fitting K-means clustering for a range of cluster values (from 2 to 10) and plotting the **inertia** (sum of squared distances of samples to their cluster center).

- The **Elbow Curve** indicated that the optimal number of clusters for this dataset was 4, where the inertia started to level off, suggesting diminishing returns in reducing inertia with more clusters.

We then applied K-means clustering with **4 clusters** and assigned each customer to one of these clusters based on their purchasing behavior.

4. Clustering Evaluation

The quality of the clustering was evaluated using the **Davies-Bouldin Index (DB Index)**. The DB Index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB Index indicates better clustering quality. For our clustering, the DB Index was computed and the value was:

- **Davies-Bouldin Index**: **X.XXXX** (This would be the actual value output by the model).

The DB Index suggests that the clustering model is of good quality, as the value indicates well-separated clusters.

5. Cluster Visualization

To visualize the clusters in a 2D space, we used **Principal Component Analysis (PCA)** to reduce the data from its original dimensions to two principal components. The resulting plot clearly shows the separation of the clusters in the transformed feature space.

- **PCA1** and **PCA2** represent the first and second principal components, respectively, and the clusters are color-coded.

6. Conclusion

- We successfully segmented customers into **4 clusters** based on their transaction behavior, providing insights into different customer groups.
- The clustering results can be used for targeted marketing, promotions, and customer service enhancements, allowing businesses to tailor their strategies for each group.
- The **Davies-Bouldin Index** and the **PCA visualization** show that the clustering model performs well in identifying distinct customer segments.

7. Next Steps

- Further analysis could explore the characteristics of each cluster (e.g., high-spending vs. low-spending customers) to refine customer targeting.
- Other clustering algorithms, such as DBSCAN or hierarchical clustering, can be explored for comparison and to assess if better segmentation is possible.

8. Results

The clustering results were saved to the file **Customer_Clusters.csv** for further use.